Supplementary tables

Supplementary table 1 Complete Heterozigosity table.

Deletions

Size*	Proportion**	Detected %***	Het %****	Det Hom %****
1	0.3	0.6	0	83
	0.4	2	0	
	0.5	10	0	
	0.6	15	0	
3	0.3	2.6	0	87
	0.4	4.6	0	
	0.5	10	0	
	0.6	21	3	
5	0.3	3	0	94
	0.4	4	0	
	0.5	13	15	
	0.6	27	15	
10	0.3	4	0	98
	0.4	14	14	
	0.5	40	20	
	0.6	68	20	
15	0.3	9	7	99
	0.4	22	41	
	0.5	66	29	
	0.6	88	30	
20	0.3	12	33	99
	0.4	38	39.7	
	0.5	69	40	
	0.6	90	45	
30	0.3	16	41.6	99
	0.4	52	89	
	0.5	85	87.5	
	0.6	97	83	
40	0.3	21	92	99
	0.4	61	89	
	0.5	87.5	93	
	0.6	98.75	92	

Insertions

Size*	Proportion**	Detected %***	Het %****	Det Hom %****
1	0.3	0.7	0	80
	0.4	7.3	0	0
	0.5	10.7	0	0
	0.6	17.3	0	0
3	0.3	3.3		86
	0.4	7.3	0	0
	0.5	9.3	0	0
	0.6	22.7	2.9	0
5	0.3	3.3	0	94
	0.4	5.3	0	0
	0.5	17.3	3.8	0
	0.6	31.3	6.4	0
10	0.3	3.3	0	99.3
	0.4	16.0	12.5	0
	0.5	28.0	14.3	0
	0.6	58.7	17.0	0
15	0.3	4.0	0	99.3

	0.4	22.0	30.3	0
	0.5	43.3	32.3	0
	0.6	71.3	34.6	0
20	0.3	6.0	44.4	99.3
	0.4	29.3	47.7	0
	0.5	57.3	47.7	0
	0.6	84.0	48.4	0
30	0.3	21.0	63.6	99.3
	0.4	71.0	91.1	0
	0.5	81.0	89.9	0
	0.6	88.7	91.0	0
40	0.3	30	88.2	99.3
	0.4	59.0	93.5	0
	0.5	88.0	96.5	0
	0.6	97.0	96.6	0

*size of the event, **proportion of sampling taken from the mutated haplotype
***recall rate for the heterozygous case ,
**** proportion of recalled indels classified as heterozygous,
*****recall rates for equivalent (same locus) Homozygous indels

Supplementary table 2: Size distributions of call by different methods and apparent specificity on Kidd dataset

size	**BreakDancer**			**SVM^2**			**Pindel**			**Valid Set**
	Total	**Valid**	**Valid %**	**Total**	**Valid**	**Valid %**	**Total**	**Valid**	**Valid %**	
1	109	18	16.51	39688	10474	26.39	180760	44286	24.50	120583
2	73	10	13.70	24120	6522	27.04	40668	11936	29.35	52706
3	97	13	10.31	20777	5572	26.82	22303	6396	28.68	19109
4	127	22	17.32	16484	4271	25.91	21854	6486	29.68	27121
5_10	5968	1331	22.30	42236	10935	25.89	18818	5493	29.19	31748
11_20	18310	4989	27.25	20470	5463	26.69	7044	1782	25.30	9533
21_30	6332	1516	23.94	5360	1272	23.73	111	18	16.22	2251
gr30	8575	1470	17.14	7396	1376	18.60				2213
										265264

Supplementary table 3: Number of calls and recall rate (Sensitivity) respect to the Kidd validation dataset (Fig 1A)

				BreakDancer	SVM^2	PinDel	Valid Set
SIZE	**BreakDancer***	**SVM^2****	**PinDel*****	**%********	**%***********	**%************	**************
1	1133	20409	44897	0.94	16.93	37.23	120583
2	1287	9444	12619	2.44	17.92	23.94	52706
3	470	4101	6504	2.46	21.46	34.04	19109
4	1032	6128	6754	3.81	22.6	24.9	27121
5_10	2246	8978	5652	7.07	28.28	17.8	31748
11_20	3070	4550	1606	32.2	47.73	16.85	9533
21_30	1131	1290	88	50.24	57.31	3.91	2251
gr30	1046	1277	54	47.27	57.7	2.44	2213
TOTAL	11415	56177	78174	4.3	21.18	29.47	265264

ST3: Validation rate (Specificity) per method per predicted size (Fig 1A) on Kidd dataset
*,**** Absolute number and proportion of indels of different size recalled by Breakdancer
,*** Absolute number and proportion of indels of different size recalled by svm2
,** Absolute number and proportion of indels of different size recalled by Breakdancer
******* Number of events in the validation set by size

Supplementary table 4: apparent specificity for each method on dbsnp and 1000 genomes data collections

Predicted_size*	**BreakDancer ****		**SVM^2*****		**PinDel*****	
	dbsnp	**1000**	**dbsnp**	**1000**	**dbsnp**	**1000**
1	51	17	79.3	70	78.2	75.5
2	58	9	79.6	69.7	83	81.6
3	54	11	79.9	69.6	81.2	87.7
4	59	20	79.3	67.9	82.8	92.3
5_10	71.2	59.5	81.1	66.8	81.9	90.9
11_20	88.6	65.5	88	65.9	79	85
21_30	87.3	62.6	86	63	68	77
gr30	72	45.8	73	46		

Proportions are respect to the numbers reported in supplementary table 2
* predicted size
** apparent specificity for BreakDancer
*** apparent specificity for SVM^2
**** apparent specificity for PinDel

Supplementary figures.

Suppl. Figure 1: Expected pattern of read mapping in the presence of different SVs

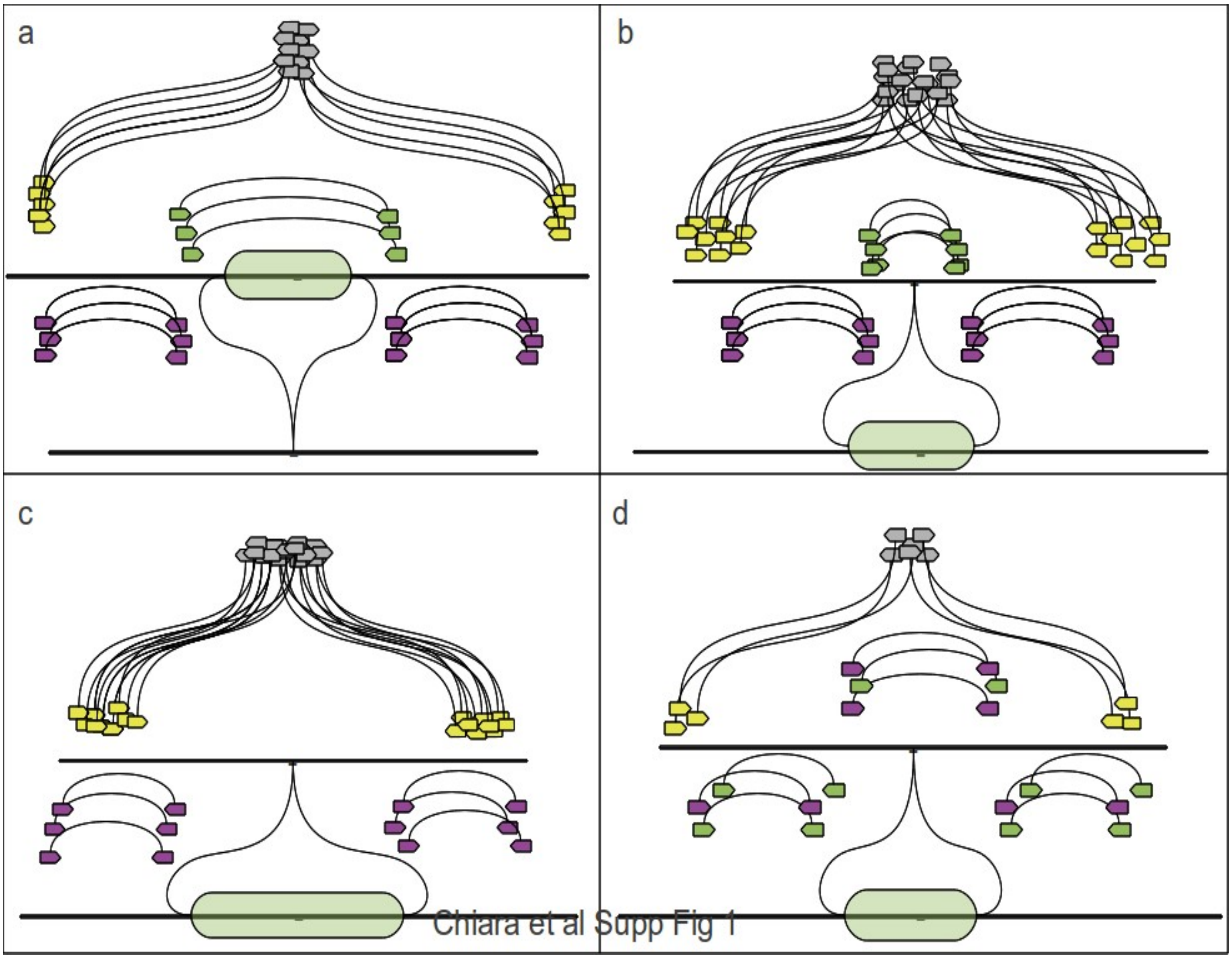


Figure S1 schematically represents the expected pattern of mapping of reads on a reference genomic sequence in the case of a deletion (a) an insertion shorter than the insert-size (b) an insertion longer than the insert-size (c) and in the presence of a particularly variable region (d). The classic approaches based on PE reads to detect indels in this scenario are: 1 define a cut-off to identify aberrant mapping reads and individuate indels as genomic clusters of aberrant mapping mates, this strategy is particularly successful for the detection of long indels; or (2) to detect smaller indels, use a particular statistic to assess whether the local insert size distribution of a particular genomic locus is significantly different from the expectations (global distribution of insert-size).

It is clear from figure 1 that in this scenario there is some additional information which could be useful to integrate in the process, as in each case an SV creates a new genomic junction, which implies that some read from the donor can't map on the reference any-more, thus generating the so called “broken-pairs”.

The difference lies in the fact that each SV event generates such broken-pairs in a specific fashion: in the case of a deletion (a) we expect a sharp peak, while for short insertion (b) we expected a broader one and eventually whence the insertion becomes too long, all we can see is a peak of broken pairs as broad as the insert-size. Furthermore, by looking at their orientation, we can distinguish between PE mapping upstream or downstream respect to an hypothetical breakpoint; this information could be used to broaden the spectrum of statistical tests used for assessing significant insert-size perturbations: indeed instead of just comparing the local distribution to the global (like others do) we could run additional test(cross-checks) by comparing upstream vs downstream, downstream vs global and upstream vs global.

Finally in figure 1 (d) illustrate show there can be some misleading signals in the case of particularly variable and localized regions, which can also lead both to the generation of peaks of broken pairs and to subtle shifts in apparent insert size distributions (although without the directional specificity observed for indels).

Suppl. Figure 2: Features used by SVM^2

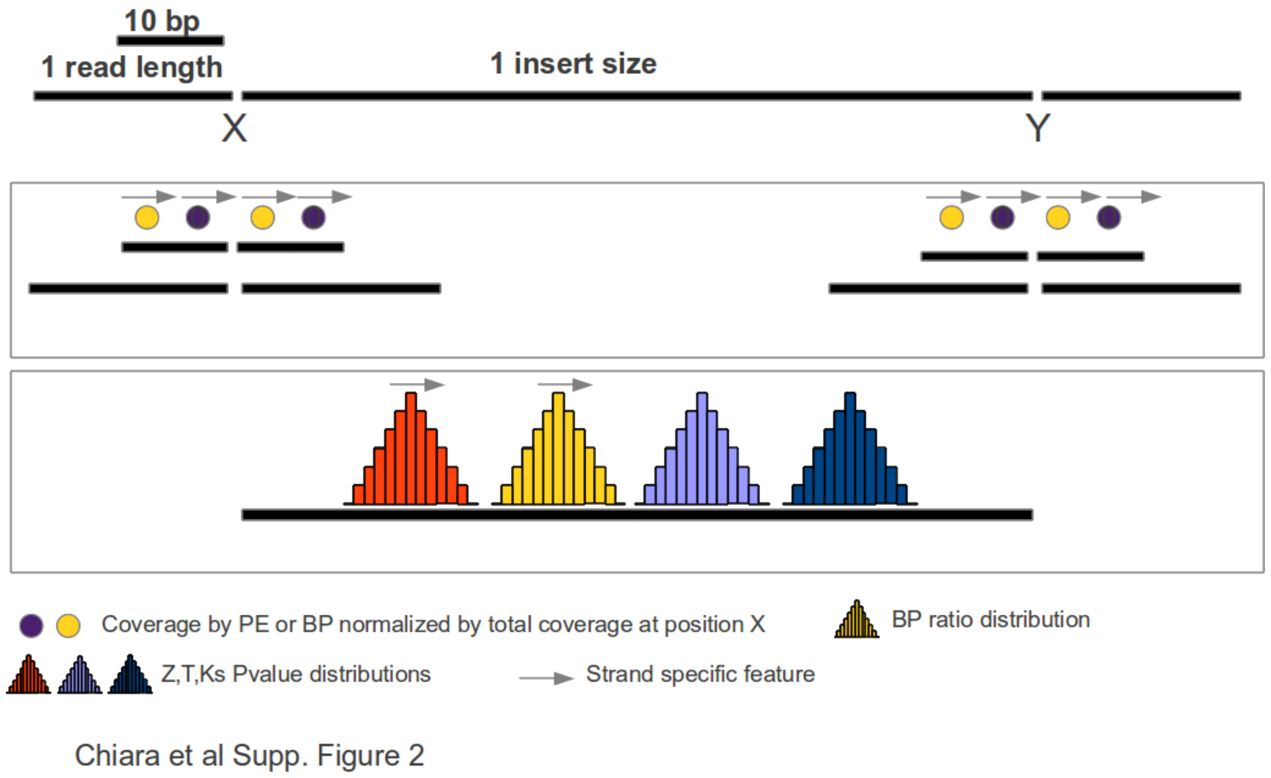

10 bp
1 read length
1 insert size
X
Y
● ● Coverage by PE or BP normalized by total coverage at position X
BP ratio distribution
Z,T,Ks Pvalue distributions
→ Strand specific feature
Chiara et al Supp. Figure 2

Figure S2: shows the localization and strandness (arrow) of the features used by our SVM. X is the position invoking the SVM, while Y is the genomic position at which mates of X are expected to be found (see methods) PE= paired end, BP=broken pairs Z= Z test, T=T-Welch test KS= Kolmogoroff Smirnov test. Features with an arrow on top are calculated on both strand All the distances are expressed in ID (see methods)

Suppl. Figure 3: Specificity by size using (3a) dbsnp130 or (3b) 1000genomes as validation set

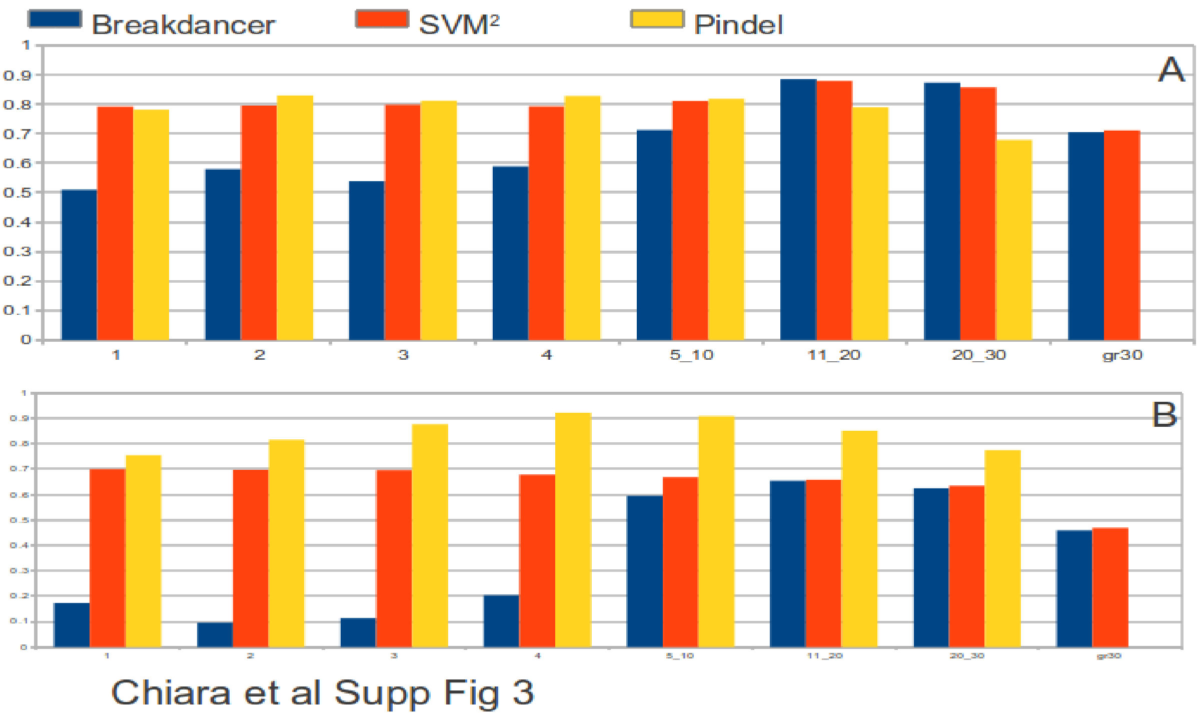
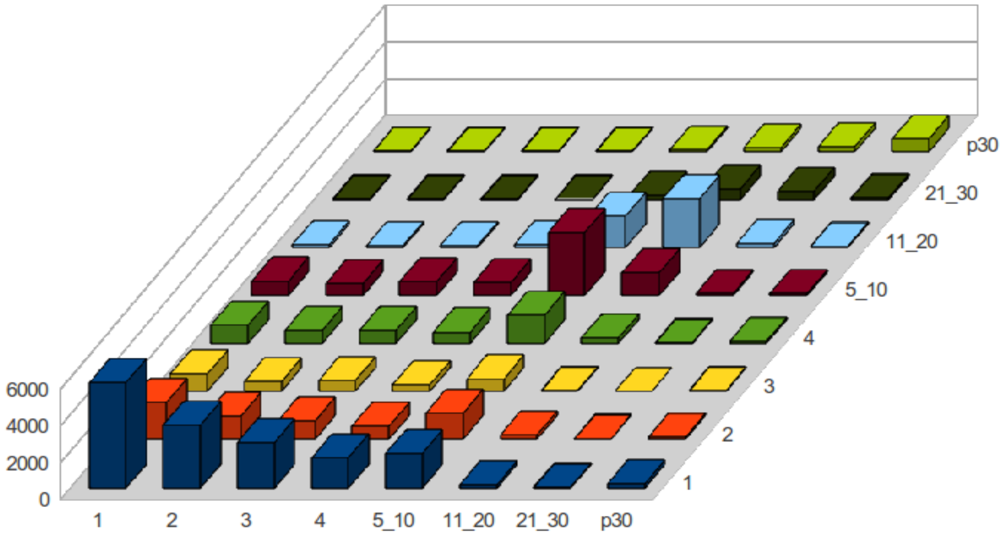


Figure S3a Equivalent to figure 1B but calculated using the whole dbsnp130 [39] as validation set
Figure S3b Equivalent to figure 1B but calculated using the entire 1000genomes SV catalog [2] as validation set

Suppl. Figure 4: 3D size distribution of predicted indels by SVM^2 validated by Kidd dataset



Chiara et al Supp. Fig 4

The X axis indicates the predicted sizes of events predicted by SVM^2 while the Z axis shows the real dimensions of the corresponding validated events from the Kidd et al. dataset. Numbers of events are shown on the Y axis.